12

Patterns Against A Background

		Page
Introduction		12-1
12A.	The invisible standard of comparison	12-1
12B.	Hanging the picture	12-3
12C.	Fitting the conventional comparison curves	12-14
12D.	Comparing distributions where the comparison curve is based on all the data	12-23
12E.	How far have we come?	12-26
	*****	
12M.	How closely does the comparison curve need to be fitted?	12-29
	*****	
12P.	IEDA processor features	12-33
	*****	
12R.	Linguistic distributions	12-34
12S.	What to do when mid-ordinates are not precise enough	12-35
12T.	Fitting comparison curves related to folded powers	12-37
	*****	
12Y.	Problems	12-44

Location of tables and plots for 12

Number	Page
1	12-4
2	12-6
3	12-7
4	12-8
5	12-10
6	12-11
7	12-12
8	12-13
9	12-16
10	12-18
11	12-20
12	12-22
13	12-24
14	12-25
15	12-27
16	12-30
17	12-32
18	12-38
19	12-39
20	12-41
21	12-42
22	12-45

12

Patterns Against A Background

The last chapter dealt with plotting distributions of counts in the case where the counts are to be plotted "against an empty sky". In doing this, we were really using a background although we did not supply one. Its use was inevitable, and unconscious rather than explicit and helpful.

In the present chapter we deal with plotting counts against a background. Here the single most important issue is the comparison of observed counts with a description of what might have been. Here we want to see a whole array of residuals along with observations and partial description.

12A. The invisible standard of comparison

From time to time we have come back to the central issue of graphical plotting:

- every picture generates a standard of comparison -- whether we want it to do so or not, whether we provide for it explicitly or try to prevent its use.

Moreover:

- if the eye can use a straight line for the standard of comparison, it will do just that.

In view of these inescapable facts, what comparison is implicitly in mind whenever we draw any of the pictures we saw in the last chapter? The strongest standard of comparison is a horizontal line (or occasionally a vertical one); when will the revealing parts of our pictures -- in this chapter usually the ends of some kind of bar -- actually follow such

a line? Exactly when the counts observed for the various intervals are all the same, when we are dealing with what is often called a rectangular distribution.

Do rectangular distributions -- or even reasonable approximations to such -- occur frequently in practice? Every honest analyst of data that I know will answer this question with a "no". Why do we use such an unlikely situation as an ideal reference? Because we are foolish? Not really. Rather because anything much better is likely to:

- have to be tailored, in some detail, to the particular distribution we are examining.

A single plotting scheme serves to make all rectangular distributions reveal themselves through horizontal straight lines. There is no way to gain an equivalent generality for any other shape with a single plotting scheme. (Proof?) As a result, if we want to go further, as we do and will, we must be prepared to make and look at pictures each of which compares a single set of data with a single particular distribution. To do this will raise a variety of new problems, but such problems are only to be expected -- once we recognize the exceptional nature of the behavior of all rectangular distributions.

Rather clearly the first general sort of distribution we shall want to look at more effectively consists of more or less symmetrical and single-humped distributions. (The example we used throughout the last chapter is a reasonable example; it comes within the scope of "more or less". We

will continue to use it.) In order to adjust our picture to be more revealing about the shape of such a distribution we need to know something about both where it is centered and how broadly it is spread.

As a consequence, we will have to give some attention to assessing location and spread and using the results of assessment to select a basis for comparison. Before we do this, however, we will begin by looking at the final step, by asking:

- given a set of values, and something with which to compare them, how do we make an effective comparison?

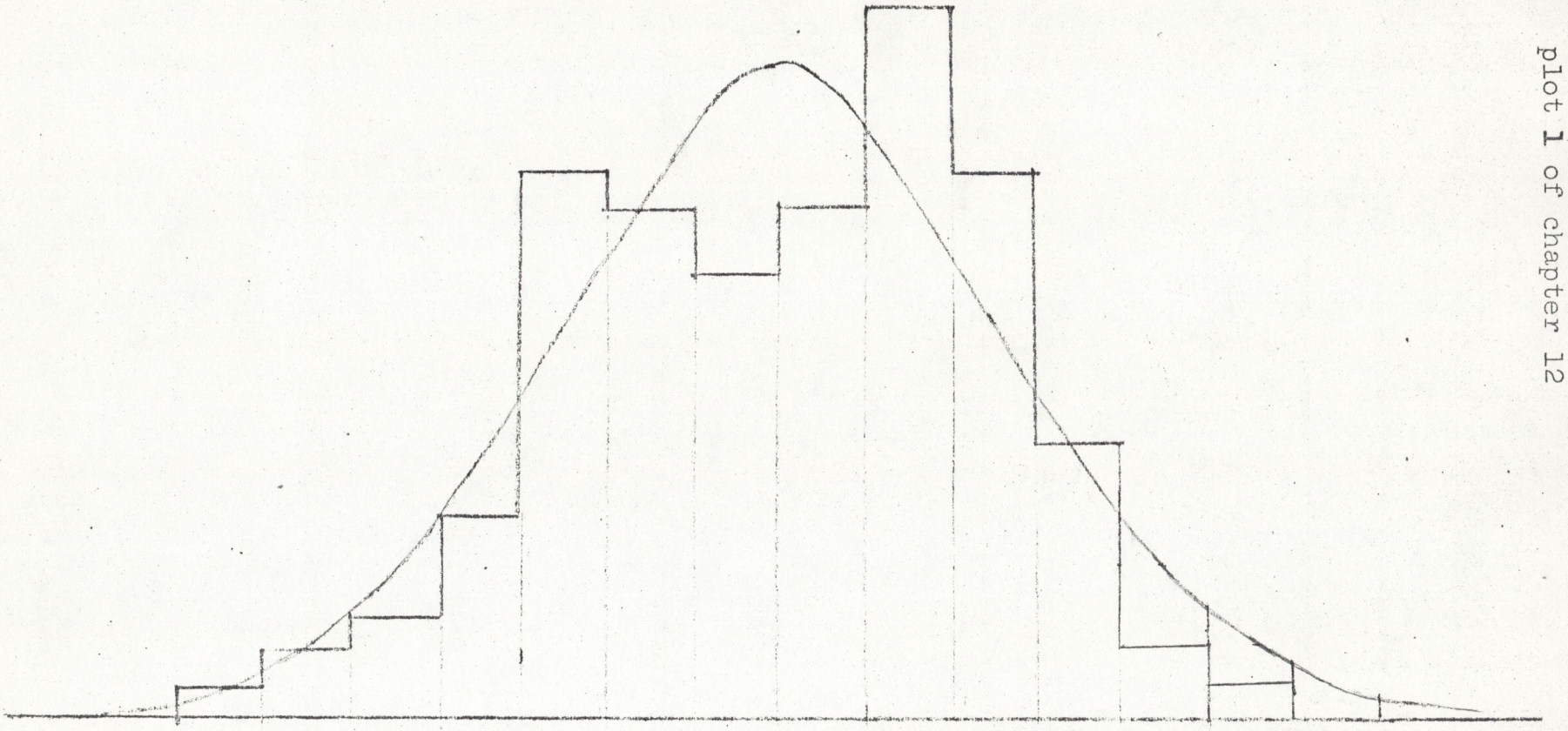
12B. Hanging the picture

Let us continue with our batch of 120 residuals. Plot 1 shows the histogram and a smooth curve of distribution, about whose origin we shall learn in the next section. In both area is proportional to count. This curve assigns, as we will also learn, respectively, 17.8, 15.4, 11.6, 7.5, 4.3, 2.1, 0.9, 0.3 and 0.1 cases to the intervals of unit length starting out from zero in either direction.

We can sit and look very hard, trying to compare the bars of the histogram with the arch of the curve. To do this is, of course, both wasteful and unwise. Comparing with a curve -- using a curve as a standard of comparison -- is always poor graphics. We should do better.

When we stop to say what we are doing, it is easy to see how to do better. We have vertical bars of specified height, which we are comparing with the vertical separation

Histogram (as in plot 5 of chapter 11) and smooth comparison curve



plot 1 of chapter 12

between a horizontal line (the x-axis) and a curve. We have been doing this by putting one end of each bar on the straight line and comparing the other end with the curve. We could make exactly equivalent comparisons by:

- putting the other end of the bar on the curve and comparing the first end with the straight line.

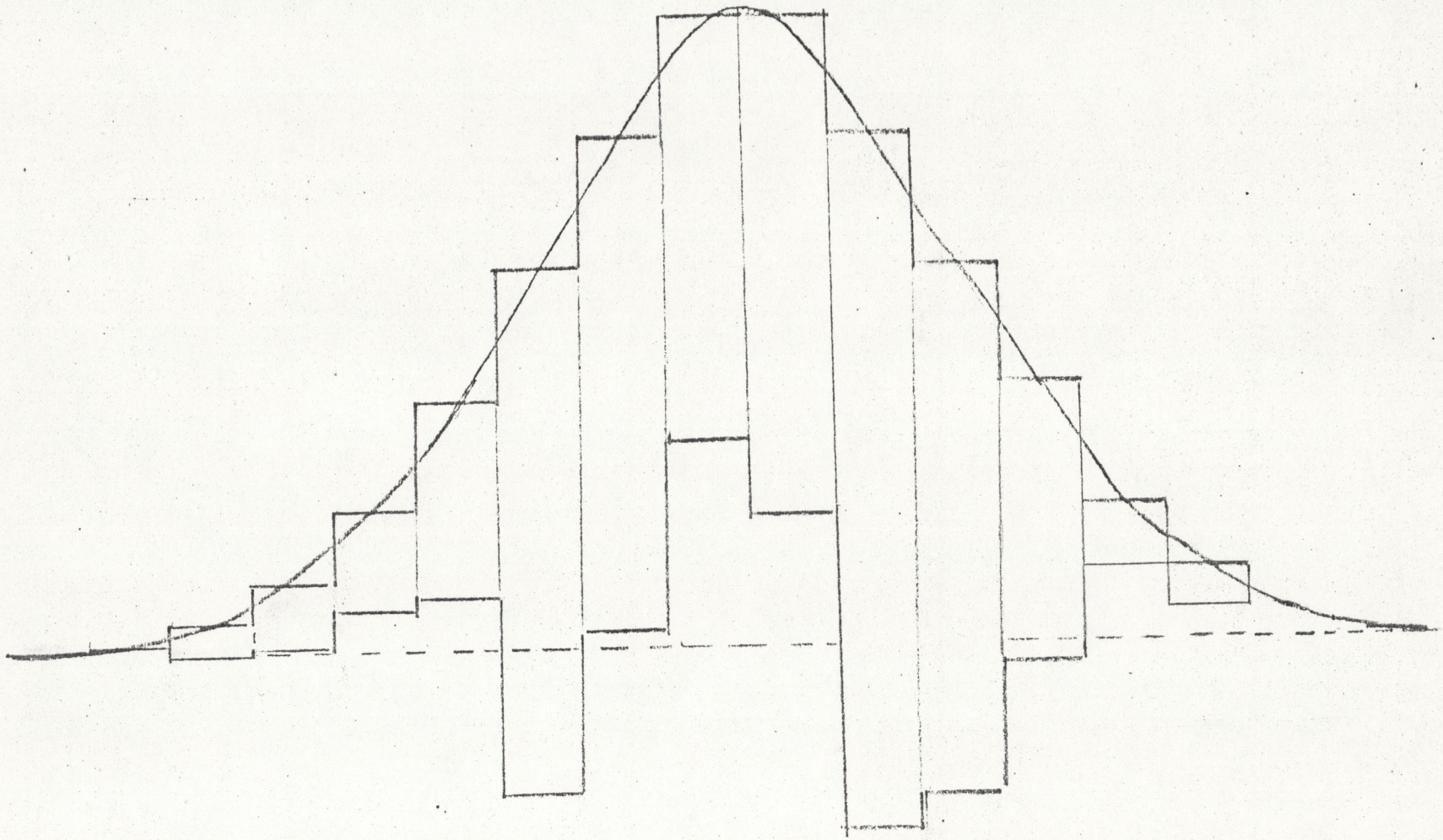
A simple change, almost a trivial one, yet a change that gains us that very important straight-line comparison standard.

Plot 2 shows the result of such a slipping. We now have a fair look at the discrepancies between batch and comparison curve. Our understanding of what is going on is, of course, hampered by the failure of the histogram to produce irregularities of about the same size all across its range.

The cure for this we know: use a rootogram -- or a rootobar plot. Let us apply this knowledge. Plot 3 shows the rootogram equivalent of plot 1. There is still great profit in slipping our bars so that they fit the curve -- and we can compare their other ends with the x-axis. Plot 4, which shows the result of such careful slipping, is the analog of plot 2, an analog that is much easier to interpret.

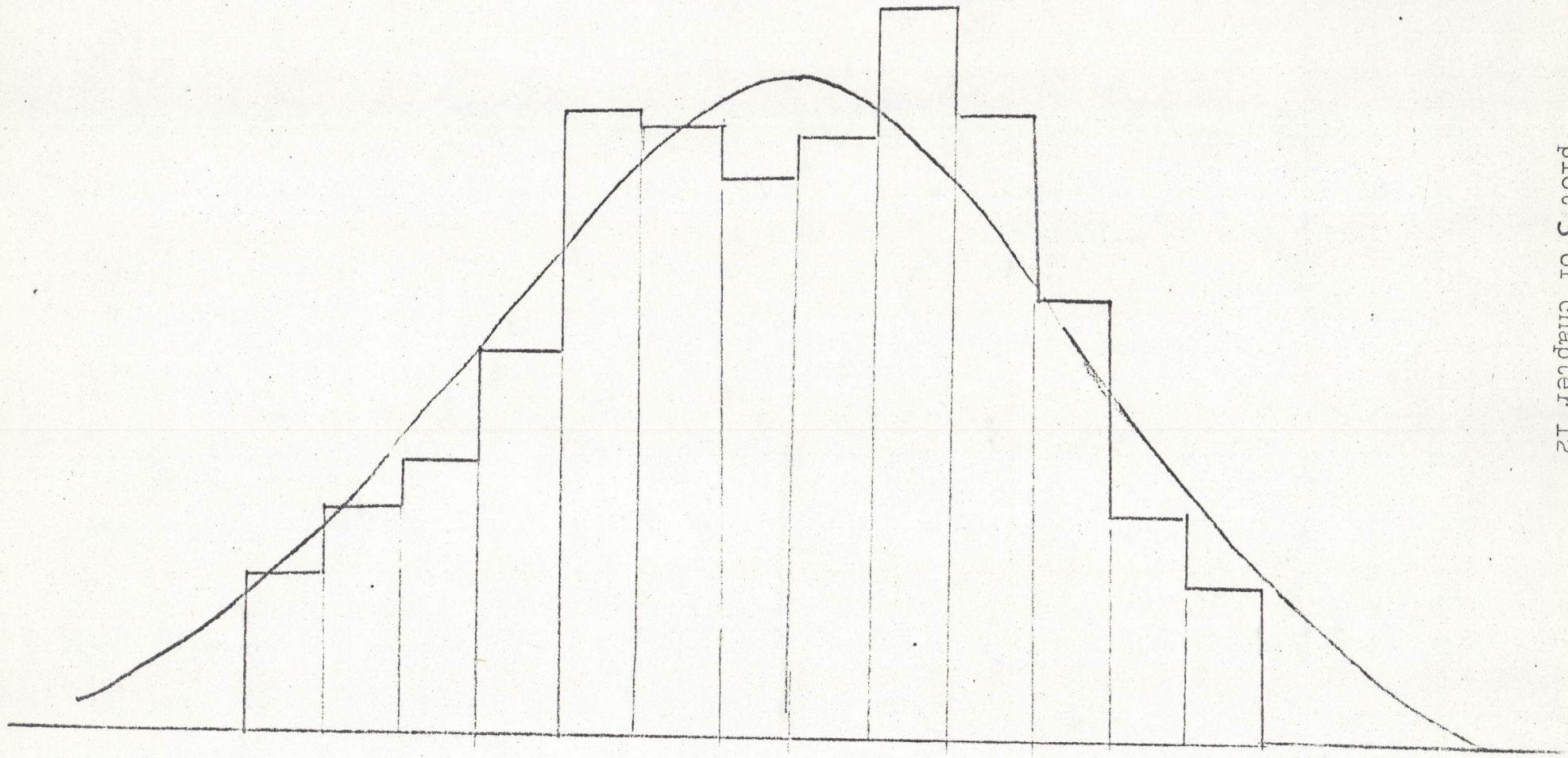
This last plot, obviously called a h a n g i n g r o o t o g r a m , provides an effective picture of useful "residuals" corresponding to the observed counts and the comparison curve. We now see quite clearly a rather recurrent ripple (almost as if periodic!) of lack of fit. Whether this

The same comparison curve with hanging histogram bars



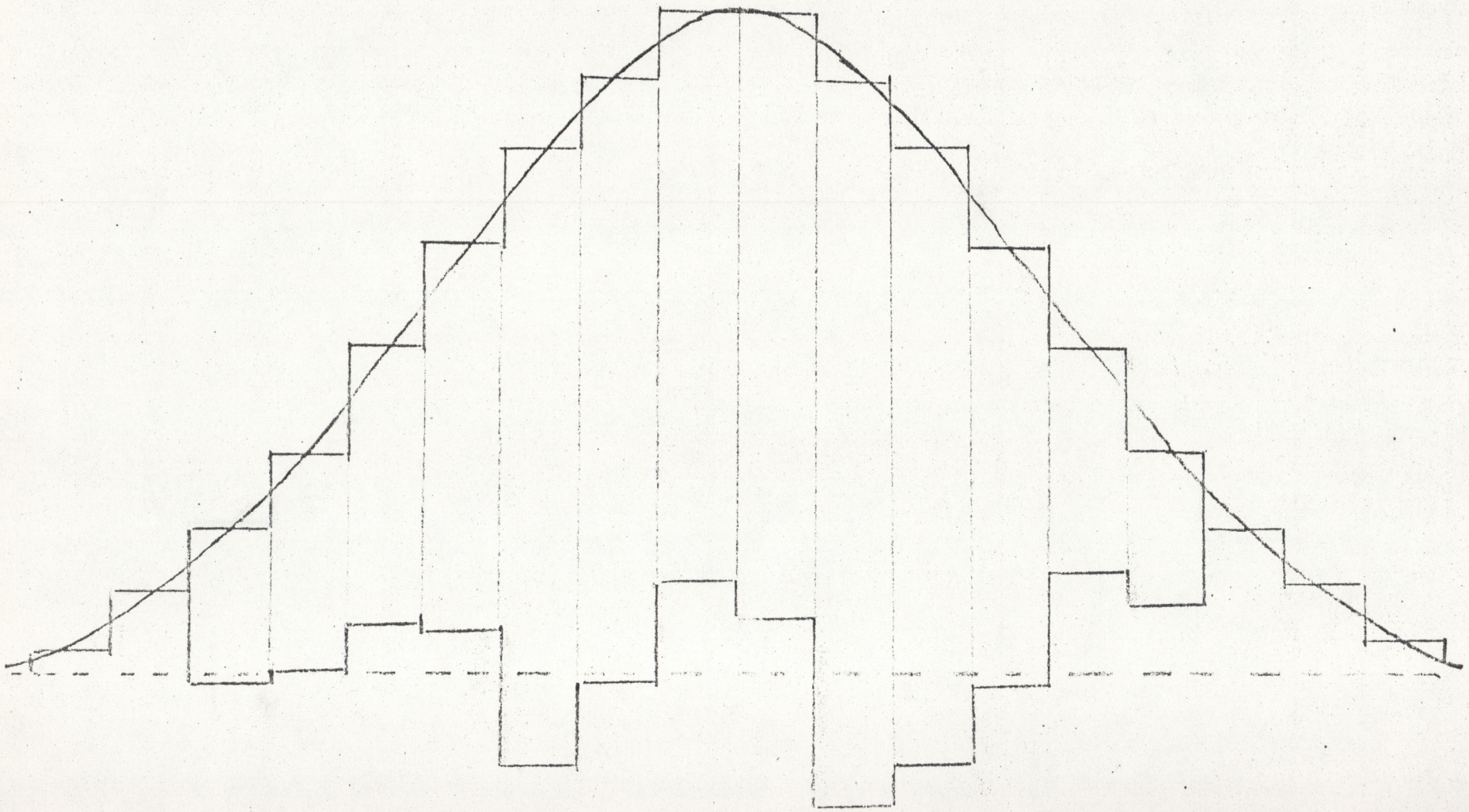
plot 2 of chapter 12

Rootogram, and the corresponding comparison curve



plot 3 of chapter 12

The same comparison curve with hanging rootogram bars.



is large enough for us to worry about is doubtful, but we at least now see it clearly.

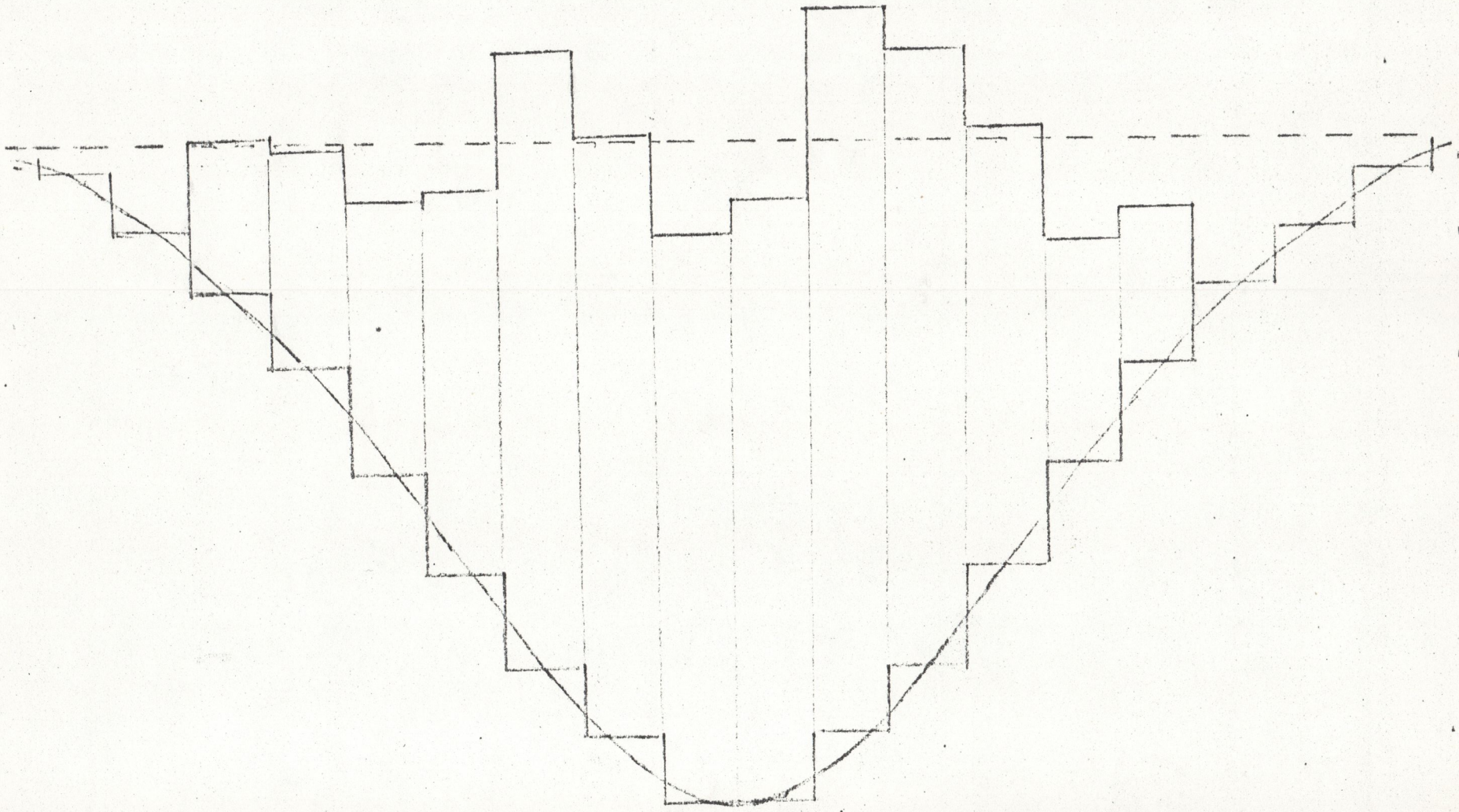
Various devices can be used to make the deviations of observed counts from the counts that would match the curve still more evident. Some like to flip the picture over as in plot 5. Such s u s p e n d e d r o o t o g r a m s are more persuasive to many who -- like the author -- are more used to look at the top ends, rather than the bottom ends, for what is essential.

We can now go further, and shade in the areas corresponding to residuals, using different directions of shading for positive and negative areas. Plot 6 shows the result. This plot shows, in ways hard to evade:

- the actual counts (on a square root scale and <u>not</u> area-true).
- the comparison curve.
- the regions of excess counts.
- regions of deficit counts.

Clearly, we can also hang or suspend rootobars. Plots 7 and 8 show the results of hanging and suspending rootobars, while keeping from the rootogram, the whole of the broken line showing the behavior of the residuals. In order to avoid an unfavorable "busy" appearance, the areas of excess and deficit in these plots have NOT been shaded.

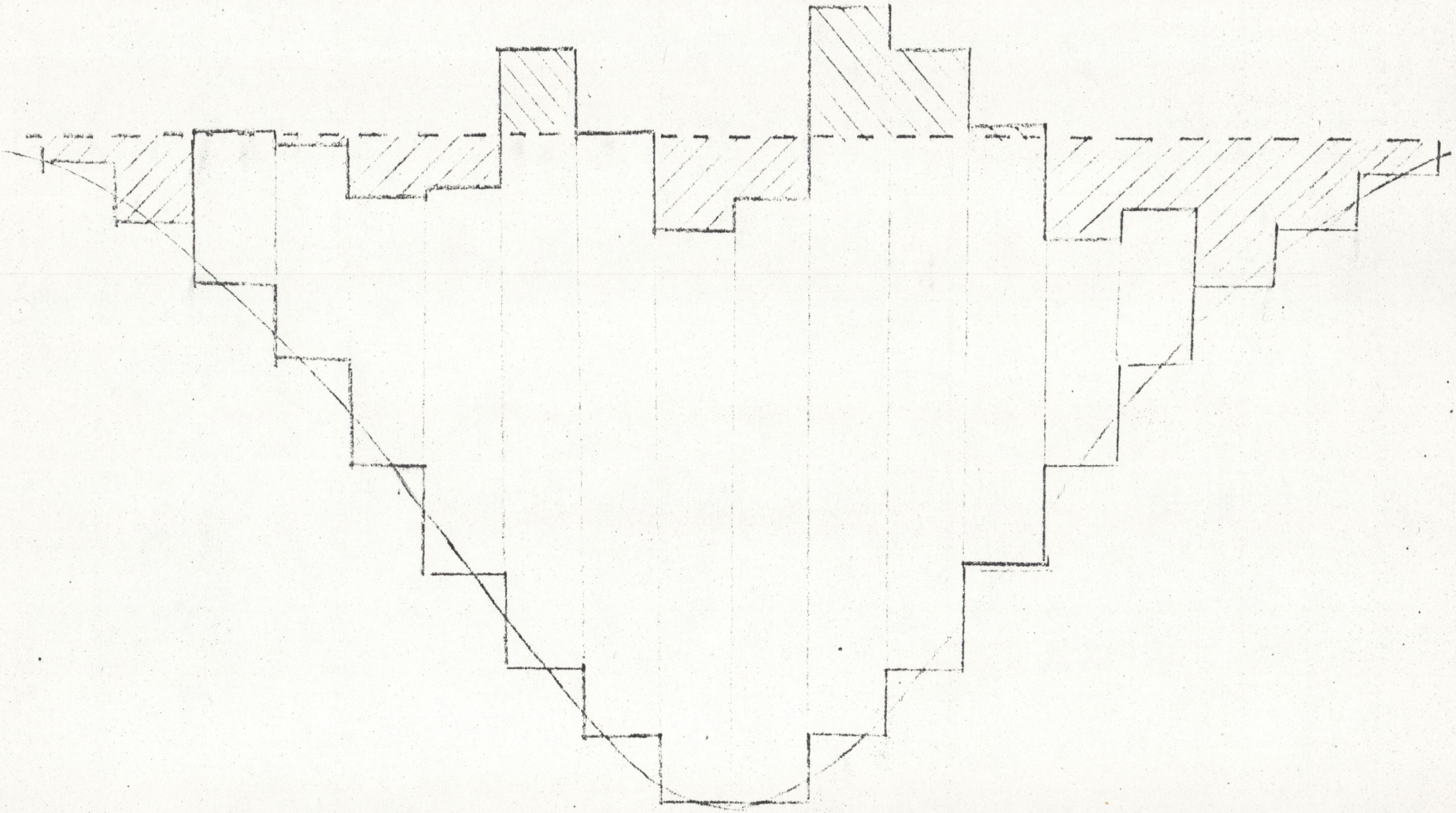
Plot 4 turned over.



plot 5 of chapter 12

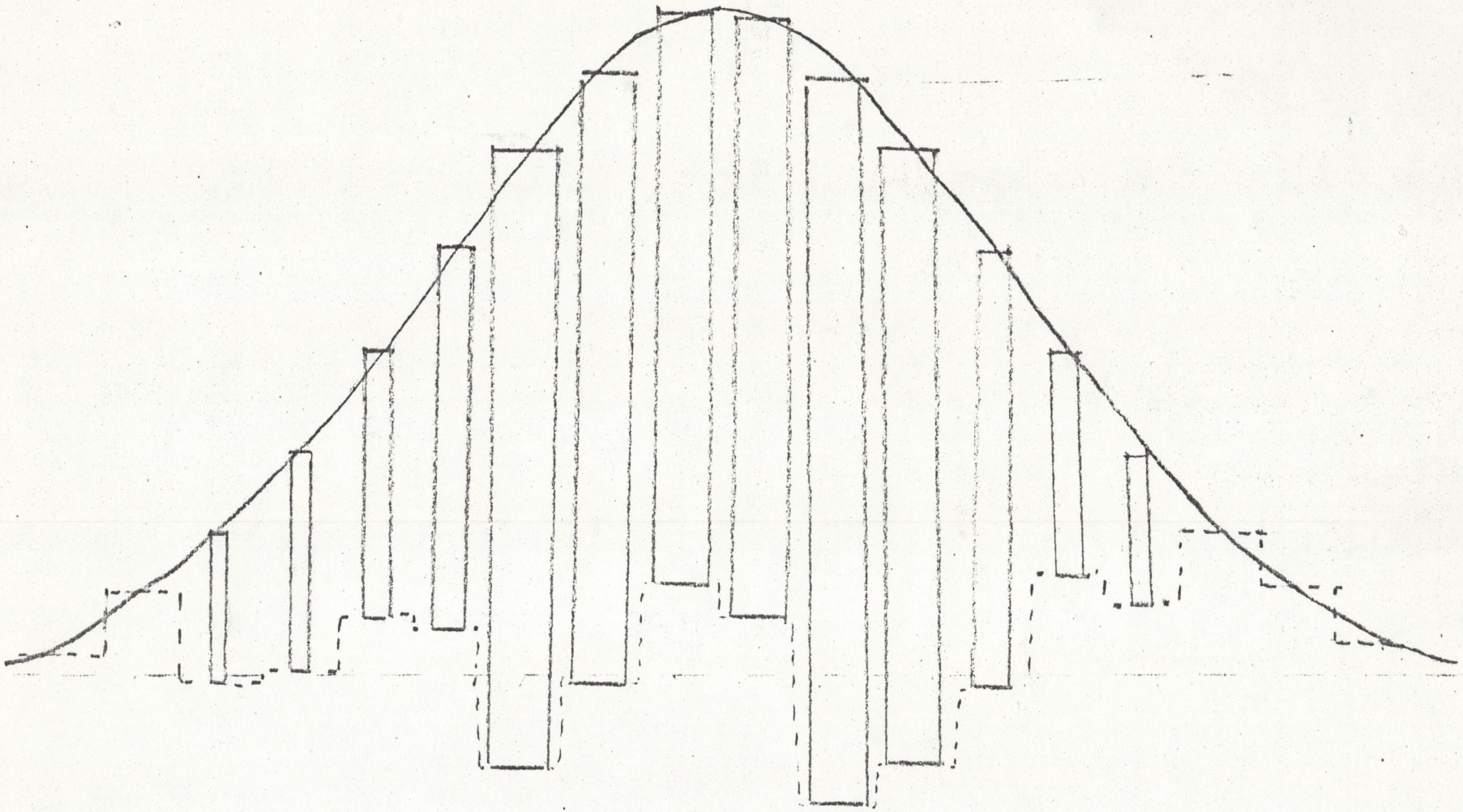
12-10

Suspended rootogram with shaded areas of excess and deficit.



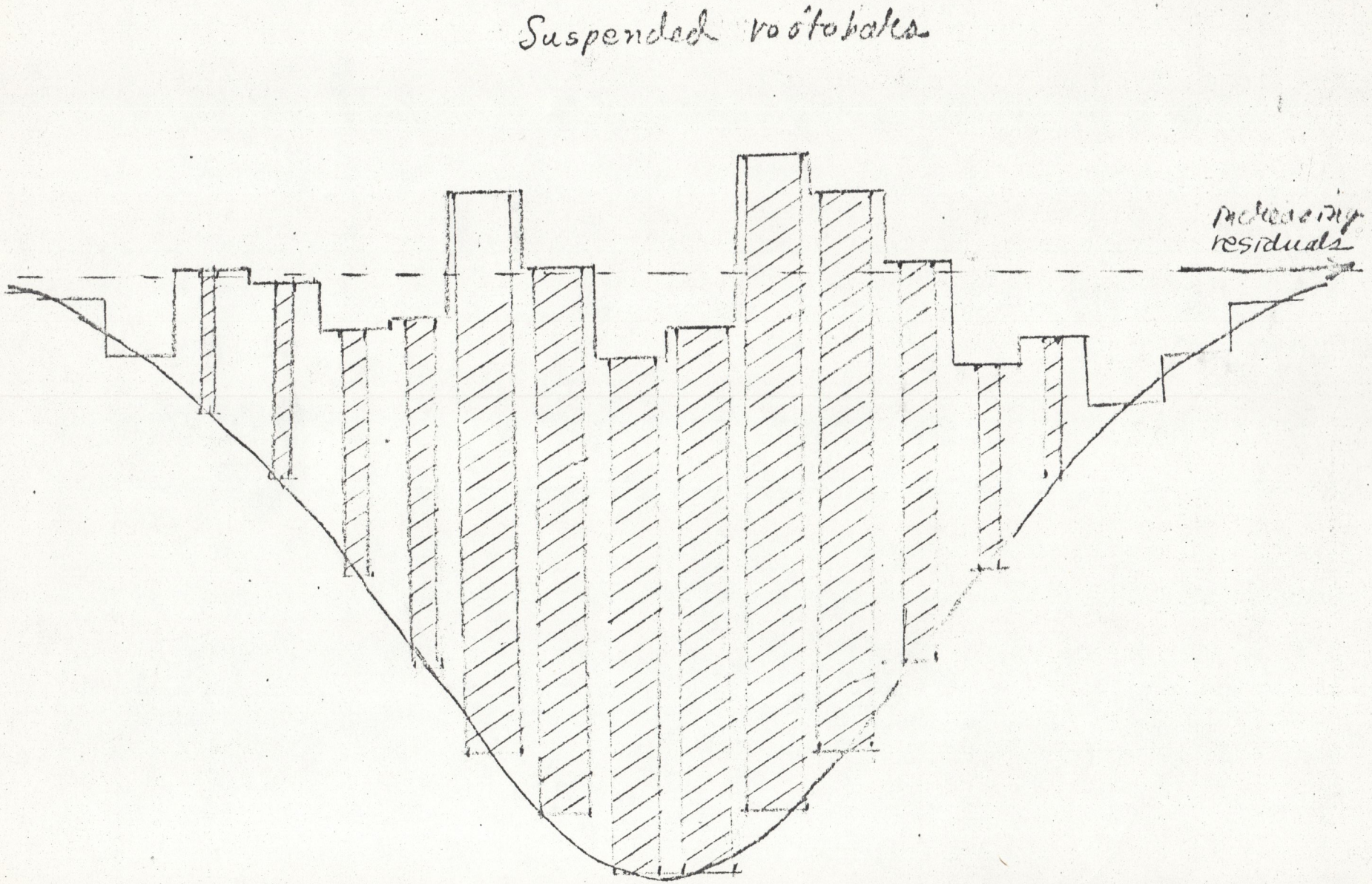
12-11
plot 6 of chapter 12

plot 7 of chapter 12



Increasing residual →

Hanging rootobars



plot 8 of chapter 12

12-13

12C. Fitting the conventional comparison curves

How are we to fit a comparison curve to a given counted distribution? Our need (as we shall later see) is for a crudely reasonable fit -- we do not need a fit of exceptionally high quality, even if there could be such.

Our fit has to start from information about the given distribution which combines information about location and information about spread. One very simple way in which to find and give this information is in terms of the upper and lower hinges. In our by-now-familiar case, display 5 of chapter 8 found these hinges to be at:

$$-1.6 \text{ and } +1.8$$

These correspond to a center at +0.1 and to a half midspread of

$$\frac{+1.8 - (-1.6)}{2} = 1.7$$

For the curves used in the last section, we decided to put the center at 0.00, just to keep things simple. Our initial intervals were each 1.00 long, which is

$$\frac{1.00}{1.7} = .59$$

of the way to the quarter point, We may as well approximate .59 by .6. The centers of our intervals, when measured in half midspreads, will now lie at

$$\tfrac{1}{2}(0.0+0.6) = 0.3$$

$$\tfrac{1}{2}(0.6+1.2) = 0.9$$

and so on -- at ...,-4.5,-3.9,-3.3,-2.7,-2.1,-1.5,-0.9,-0.3, +0.3,+0.9,+1.5,+2.1,+2.7,+3.3,+3.9,... .

Our next problem is to calculate how many counts the comparison curve assigns to each of our intervals -- in particular, say, to the interval from +0.6 to +1.2. For graphical accuracy, it will usually be precise enough to take this count, given by the area under this part of the curve, as proportional to the height of the curve at the center of the interval. (For our selected interval, the height at +0.9.) Thus we need:

- to find the r e l a t i v e heights of the curve at the midpoints of our intervals.
- to scale up -- or down -- these heights by a factor chosen to make the total count equal to -- or close to -- the 120 which is the observed count.

The basis for these calculations is table 9, which gives both heights and, where likely to be useful, remaining fractions for three quite different comparison curves: rectangular, Gaussian, and Cauchy. The middle one is often miscalled "normal" -- miscalled, since it is not usual for data to follow such a distribution with high precision. Roughly speaking, Gaussian distributions only exist in abstract theory. But this absence of Gaussian distributions from the

table 9 of chapter 12

Relative Heights for Three Comparison Curves

(Each scaled to put the hinges at ±1.0)

Scale	Rect.	Gauss	Cauchy
0.0	1.00	1.00	1.00
±0.1	1.00	1.00	.99
±0.2	1.00	.99	.96
±0.3	1.00	.98	.92
±0.4	1.00	.96	.86
±0.5	1.00	.94	.80
±0.6	1.00	.92	.74
±0.7	1.00	.89	.67
±0.8	1.00	.86	.61
±0.9	1.00	.83	.55
±1.0	1.00	.80	.50
±1.1	1.00	.76	.45
±1.2	1.00	.72	.41
±1.3	1.00	.68	.37
±1.4	1.00	.64	.34
±1.5	1.00	.60	.31
±1.6	1.00	.56	.28
±1.7	1.00	.52	.26
±1.8	1.00	.48	.24
±1.9	1.00	.44	.22
±2.0	falls	.40	.200
±2.1	0.00	.37	.185
±2.2	.00	.33	.171
±2.3	.00	.30	.159
±2.4	.00	.27	.148
±2.5	.00	.24	.138
±2.6	.00	.22	.129
±2.7	.00	.194	.121
±2.8	.00	.121	.113
±2.9	.00	.150	.106
±3.0	.00	.131	.100
(*)	(.000)	(.022)	(.102)

Scale	Rect. (*)	Gauss (*)	Cauchy (*)
3.0	.00	.131	.100
	(.000)	(.022)	(.102)
3.2	.00	.100	.089
	(.000)	(.016)	(.097)
3.4	.00	.074	.080
	(.000)	(.011)	(.091)
3.6	.00	.054	.072
	(.000)	(.008)	(.086)
3.8	.00	.038	.065
	(.000)	(.005)	(.082)
4.0	.00	.027	.059
	(.000)	(.004)	(.078)
4.2	.00	.0185	.054
	(.000)	(.002)	(.074)
4.4	.00	.0125	.049
	(.000)	(.002)	(.071)
4.6	.00	.0084	.045
	(.000)	(.001)	(.068)
4.8	.00	.0054	.042
	(.000)	(.001)	(.066)
5.0	.00	.0035	.038
	(.000)	(.000)	(.063)
6.0	.00	.0002	.027
	(.000)	(.000)	(.052)
7.0	.00	.0000	.020
	(.000)	(.000)	(.045)
8.0	.00	.0000	.0154
	(.000)	(.000)	(.040)
9.0	.00	.0000	.0122
	(.000)	(.000)	(.036)
10.0	.00	.0000	.0099
		(.000)	(.034)

(*) These entries give the fraction of area in a single tail beyond the scale value for the line immediately above -- the one-tail area.

real world does nothing to keep Gaussian distributions from serving us reasonably well:

- as a convenient and conventional basis for comparison, to be used both with data and with theoretically defined distributions.

- as a first approximation, when we know only -- or wish to use only the knowledge of -- something about location and spread and yet require a more detailed distribution for what we are to do.

Our present activity, making hanging rootograms, is a special case of each of these uses. We are involved in comparison -- and we do want to use only information about location and spread. Thus it should be no surprise that most hanging histograms -- at least on the first try -- are hung on Gaussian comparison curves.

Before we go on with our example, we ought to strengthen our grasp on the three shapes of comparison curve to which table 9 applies. What are these shapes like? Plot 10 shows the three shapes. It it we have chosen to match among the curves:

- the spacing (and location) of the hinges (quartiles) (shown by the beginning of shading).

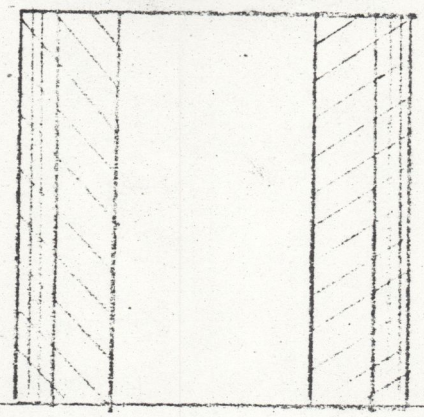
- the central height.

Notice carefully what it means for one shape of distribution to have "longer tails" than another. This is NOT just a matter of breadth. It is easy to think of a very broad

plot **10** of chapter 12

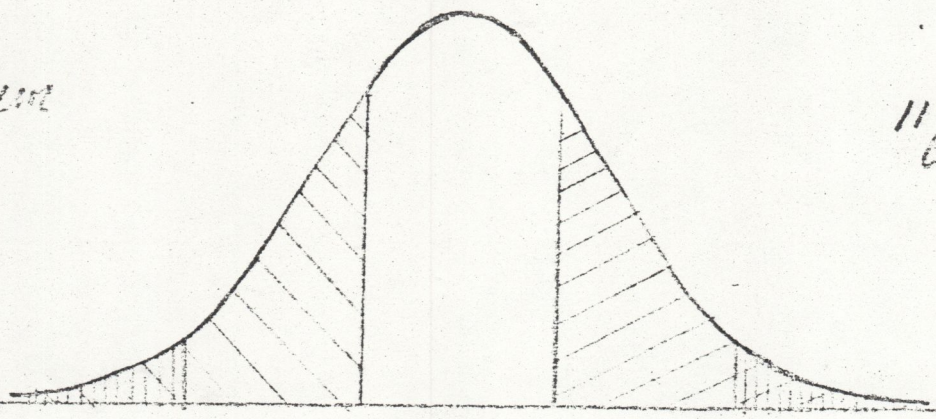
The Three Shapes of Comparison Curves Dealt with in table 9

Short tails



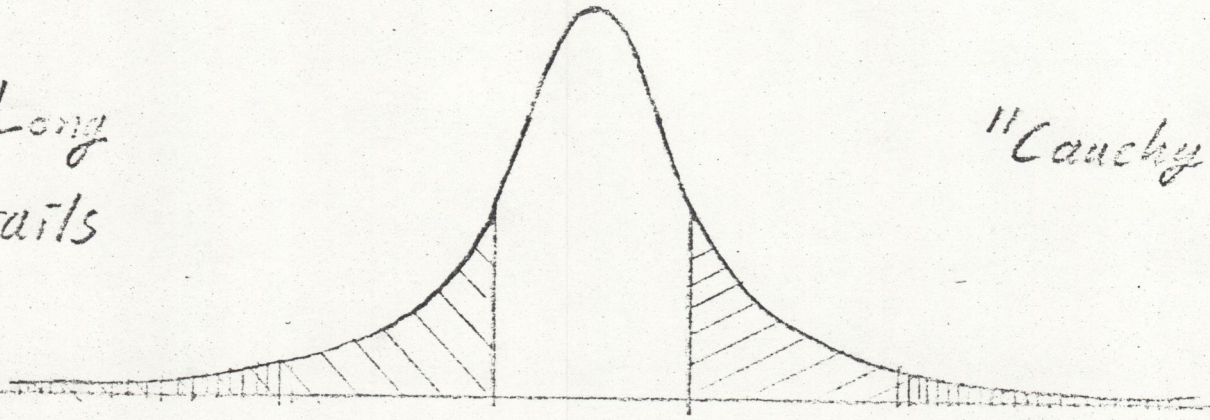
"Rectangular"

Medium tails



"Gaussian"

Long tails



"Cauchy"

Note: Middle 50% unshaded, both end 10%'s doubly shaded

Gaussian distribution being compared with a very concentrated Cauchy distribution -- when the latter will still be said to have longer tails.

"Long tails" can be fairly well described as a matter of how far the extreme tails spread in comparison with what would be reasonably anticipated in view of the spread of the middle part of the distribution -- perhaps the middle 50%, perhaps the middle 90%.

The relation of behavior to such anticipation is quite different for our three distributions.

Turning back to our example, then, we need to carry out the calculations set out in table 11. Here we use the relative height at the center of the interval to determine, to a sufficient approximation, how many cases out of 120 the comparison curve would assign to each interval.

To calculate the correct scaling factor, we need:

- first, to allow for the part of the area under the comparison curve not included in any of the chosen intervals by converting this part into counts. (Here found, using table 9, to be .004 of 120 or 0.5 cases.)
- second, to divide the remainder, here 119.5 out of 120, by the sum of the relative heights in order to find the scaling factor.

Notice that the scaled sizes of the intervals with few counts have been carried to more decimals in order that their square roots be well defined.

The last three columns of square roots give the numbers (that should have been) plotted in plots 3 through 8. (Note minor changes in areas calculated for the curve.)

table 11 of chapter 12

Calculation of a Gaussian Comparison Curve
and its Relation to the Data of display 5 of chapter 8
(All intervals of equal length)

			------Square roots------		
Center*	Height**	Scaled***	Scaled	Obs'd	Diff.
-3.9	.032	.62	.79	1.0	+.2
-3.3	.087	1.68	1.30	1.4	+.1
-2.7	.194	3.74	1.94	1.7	-.25
-2.1	.37	7.2	2.69	2.45	-.25
-1.5	.60	11.6	3.40	4.0	+.6
-0.9	.83	16.0	4.00	3.9	-.1
-0.3	.98	19.0	4.36	3.6	-.75
+0.3	.98	19.0	4.36	3.9	-.45
+0.9	.83	16.0	4.00	4.7	+.7
+1.5	.60	11.6	3.40	4.0	+.6
+2.1	.37	7.2	2.69	2.8	+.1
+2.7	.194	3.74	1.94	1.4	-.55
+3.3	.087	1.68	1.30	1.0	-.2
+3.9	.032	.62	.79	0.0	-.8
	Sum 5426	119.68			
	76				
	6.186				

* Center of interval, in units of half midspread.

** Relative height of curve.

*** Heights multiplied by 19.35, (the scaling factor, calculated below)

Beyond +4.2 .002
Beyond -4.2 .002
.004

(Reduces 120 to 119.5)

$\frac{119.5}{6.186} = 19.35 =$ scaling factor

If we halve the intervals for positive values, so that the widths are no longer all alike, we have to add one step to the calculation. Table 12 shows the result.

table 12 of chapter 12

Calculation of a Gaussian Comparison Curve
With Unequal Intervals

Center	Widths*		Height	Prod**	Scaled***		Square roots Scaled	Obs'd****	Diff.
-3.9	0.6	2	.032	.064	.31	.62	.56	0.7	+1.5
-3.3	0.6	2	.087	.174	.84	1.68	.92	1.0	+0.1
-2.7	0.6	2	.194	.388	1.85	3.70	1.36	1.2	-.15
-2.1	0.6	2	.37	.74	3.6	7.1	1.90	1.7	-.2
-1.5	0.6	2	.60	1.20	5.8	11.6	2.40	2.8	+.4
-0.9	0.6	2	.83	1.66	8.0	16.0	2.74	2.75	.0
-0.3	0.6	2	.98	1.96	9.5	18.9	3.08	2.5	-.6
+0.15	0.3	1	.99	.99	9.6	9.6	3.10	2.8	-0.3
+0.45	0.3	1	.95	.95	9.2	9.2	3.03	2.65	-.4
+0.75	0.3	1	.88	.88	8.5	8.5	2.92	3.9	+1.0
+1.05	0.3	1	.78	.78	7.5	7.5	2.74	2.65	-0.1
+1.35	0.3	1	.66	.66	6.4	6.4	2.52	3.5	+1.0
+1.65	0.3	1	.55	.55	5.3	5.3	2.30	2.0	-0.3
+1.95	0.3	1	.42	.42	4.1	4.1	2.02	2.25	+.25
+2.25	0.3	1	.32	.32	3.1	3.1	1.76	1.7	-.05
+2.55	0.3	1	.21	.21	2.0	2.0	1.41	1.4	.0
+2.85	0.3	1	.160	.160	1.53	1.53	1.23	0.0	-1.2
+3.15	0.3	1	.107	.107	1.01	1.01	1.00	1.0	0.0
+3.45	0.3	1	.069	.069	.66	.66	.81	0.0	-0.8
+3.75	0.3	1	.042	.042	.41	.41	.64	-	-0.6
				8334		8121			
				393		377			
				12.324		118.91			

*Actual and relative widths.

**Product of height and width.

***Scale factor x height (gives counts/unit base) and scale factor x Prod. (gives counts/interval).

****Square roots of observed count/width.

Calculation of Scaling factor { Beyond -4.2 .002
Beyond +3.9 .005
.007x120 = 0.8

$\frac{119.2}{12.324} = 9.67$

12D. Comparing distributions where the comparison curve is based on all the data

In a fair number of situations, the best comparison curve to use when comparing different parts of a body of data is that furnished by all the data taken together. In 1951, H. S. Darling reported on the pink bollworm, *Platyedra Gossypiella*, as a pest of cotton at Zeidab in the Northern Sudan. In particular, he reported on seven years of experiments in each of which 10,000 bolls of cotton were picked at Zeidab in late November and early December, and caged under conditions where the number of pink bollworm moths emerging could be counted month by month. The counts -- and their square roots -- are given in table 13, together with means over seven years and the differences from these means.

The corresponding suspended rootograms for three selected years -- 1940, 1942, 1948 -- are shown in plot 14. Here the basic distribution, from which the square roots are suspended, is the mean over all seven years of the square roots for the counts of moths emerging in a given month. Each of these pictures shows its own distinct character quite clearly. 1948 shows a substantial surplus emergence for the first nine months of the year. 1942 shows very high emergence in January, followed by smaller, more or less oscillatory, deficits and surpluses. 1940 shows a tendency to a strong deficit from (April or) May onward. It is very hard to see as much from any table of numbers as we can easily see from these pictures.

table 13 of chapter 12

Numbers of Pink Bollworm Moths Emerging Each Month
From 10,000 Bolls of Cotton

a) COUNTS

Month	1940	1941	1942	1943	1944	(*)	1948	1949
Jan	96	68	285	100	40		203	12
Feb	57	39	44	74	96		89	8
Mar	71	47	25	20	46		116	18
Apr	38	8	47	60	20		79	21
May	0	0	28	52	16		42	29
June	0	0	10	13	1		14	0
July	0	0	1	18	2		7	0
Aug	0	0	4	19	1		18	4
Sept	0	0	0	2	0		3	1
Oct	0	0	0	1	0		0	0

b) SQUARE ROOTS (values of $\sqrt{\text{count}} + \sqrt{1+\text{count}}$, from Table 2 of Chapter 10):

								(**)
Jan	19.6	16.6	33.8	20.0	12.7	28.5	7.1	(19.7)
Feb	15.2	12.6	13.3	17.3	19.6	18.9	5.8	(16.1)
Mar	16.9	13.8	10.1	9.0	13.6	21.6	8.6	(13.4)
Apr	12.4	5.8	13.8	15.6	9.0	17.8	9.2	(13.9)
May	1.0	1.0	10.6	14.5	8.1	13.0	10.8	(8.4)
June	1.0	1.0	6.5	7.3	2.4	7.6	1.0	(3.8)
July	1.0	1.0	2.4	8.6	3.1	5.5	1.0	(3.2)
Aug	1.0	1.0	4.2	8.8	2.4	8.6	4.2	(3.1)
Sept	1.0	1.0	1.0	3.1	1.0	3.7	2.4	(2.0)
Oct	1.0	1.0	1.0	2.4	1.0	1.0	1.0	(1.2)

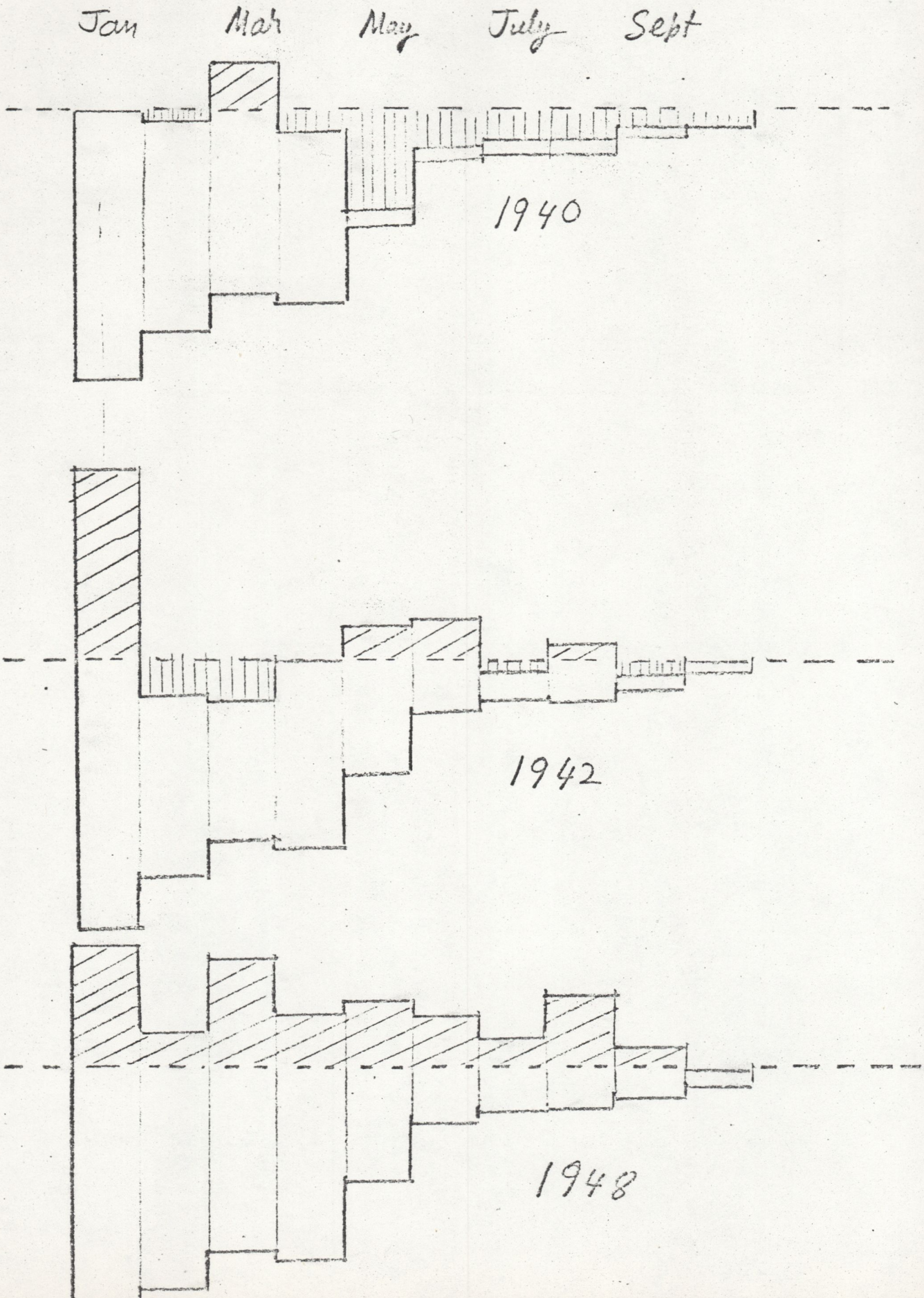
c) DIFFERENCES (of square roots from monthly means):

Jan	-0.1	-3.1	14.1	0.3	-7.0	8.8	-12.8
Feb	-0.9	-3.5	-2.8	1.2	3.6	2.8	-10.3
Mar	3.5	0.4	-3.3	-4.4	0.2	8.2	-4.8
Apr	-1.5	-8.1	-0.1	-1.7	-4.9	3.9	-4.7
May	-7.4	-7.4	2.2	6.1	-0.3	4.6	2.4
June	-2.8	-2.6	2.7	3.5	-1.4	3.8	-2.8
July	-2.2	-2.2	-0.8	5.4	-0.1	2.3	-2.2
Aug	-2.1	-2.1	1.1	5.7	-0.7	5.5	1.1
Sept	-1.0	-1.0	-1.0	1.1	-1.0	1.7	0.4
Oct	-0.2	-0.2	-0.2	1.2	-0.2	-0.2	-0.2

(*) Work interrupted by locust campaigns.
(**) Mean over seven years.

Source: H. S. Darling 1951. Pink Bollworm, Platyedra Gossypiella (Saund.), as a pest of cotton at Zeidab, Northern Sudan. Bull. Entomolog. Res. 42: 157-167.

Distribution of pink bollworm moths by month of emergence. (Square-root scale; base 12 mean square-root for seven years, 1940-1949)



A different format, probably rather clearer, is shown in plot 15, which repeats two of the years from plot 14.

12E. How far have we come?

This chapter has dealt with the display of patterns of counts -- distributions of values -- in explicit comparison with a reference pattern, often a smooth comparison curve. We began by noticing that the displays of the last chapter -- because viewers are as they are -- involve an implicit comparison. Even though no trace of its presence be on the paper, a rectangular distribution is implicitly present in such displays.

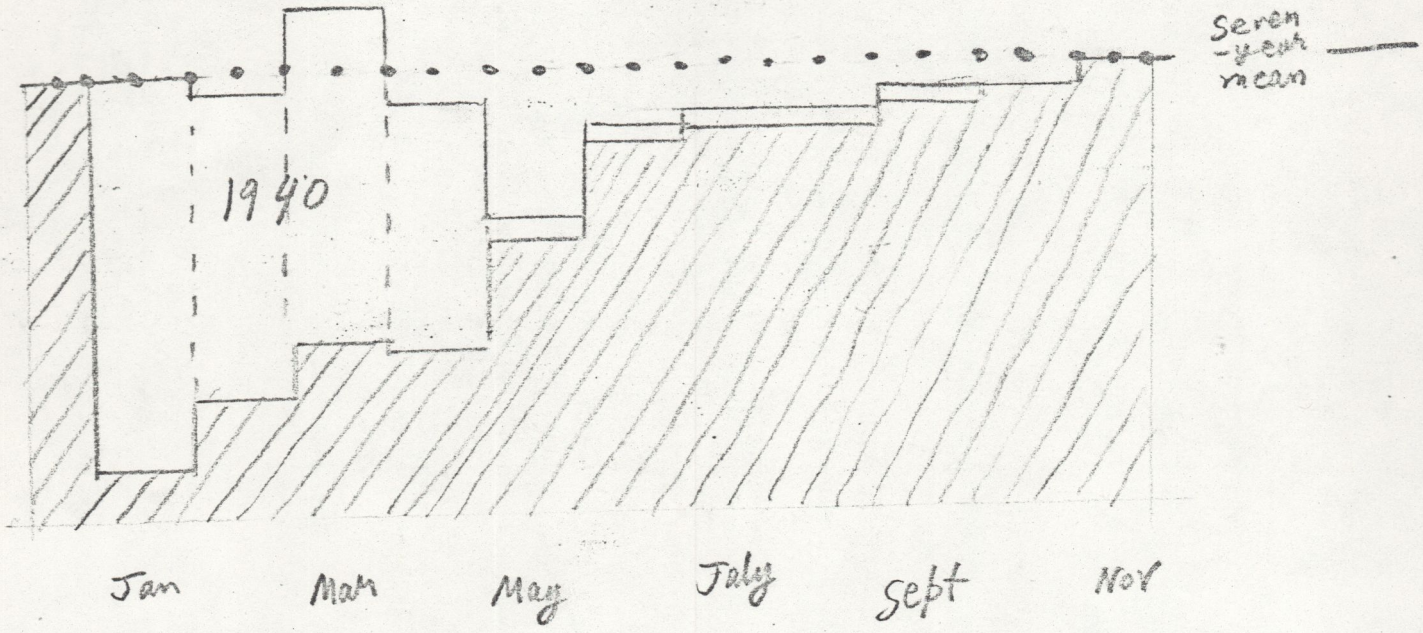
For most patterns of counts per unit interval, a curve standard of reference is better. It needs to be explicit, since there are so many different curves. Its use is helpful even when we can see that a different curve would be more clearly matched by the data.

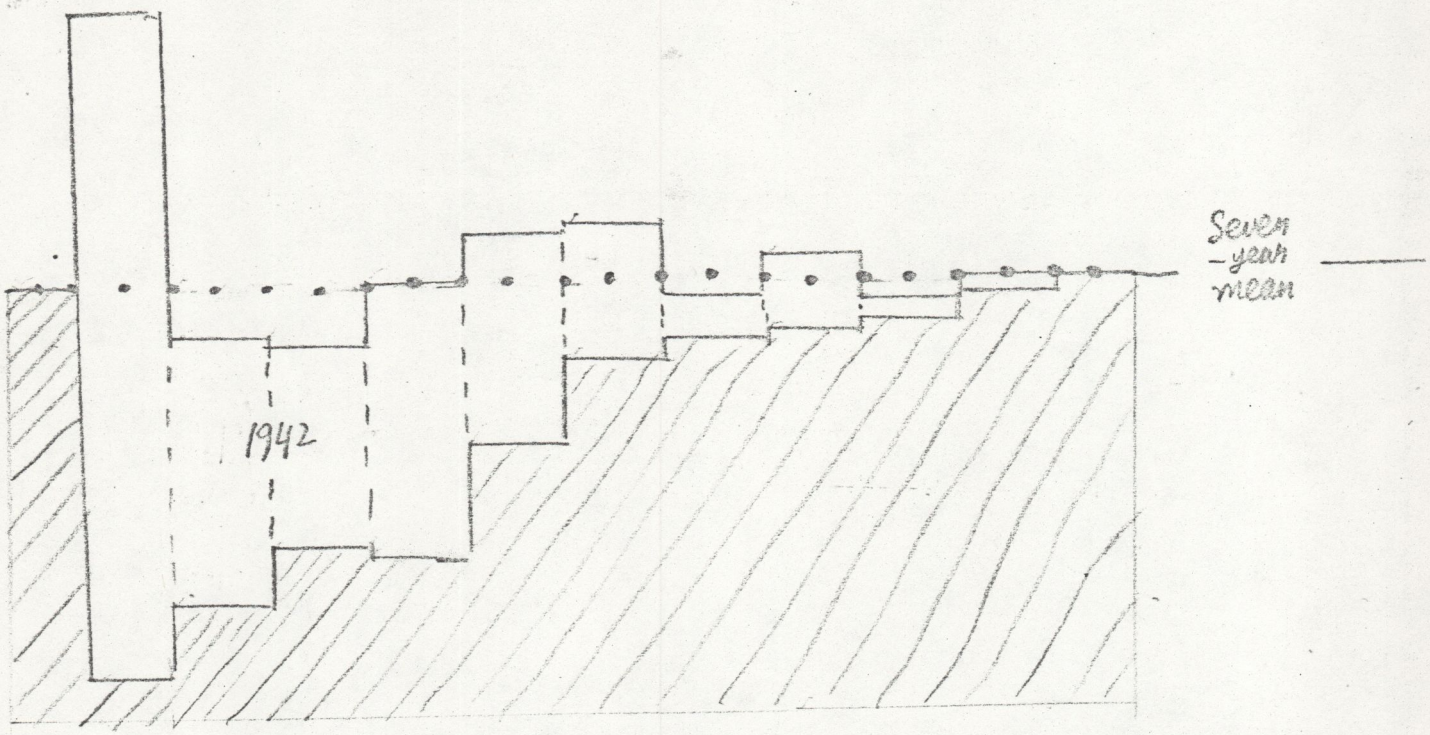
Comparisons, of points or of ends of bars, with straight lines are better than with curves. Accordingly, we have learned to hang or suspend the bars of our histograms, rootograms or rootobars.

Some of the advantage of rootobars remains when they are hung or suspended, but the great advantage of the area principle seems to be confined to those situations where there is no other explicit standard of comparison. Accordingly, hanging or suspended r o o t o g r a m s seem to be the display of choice, with hanging or suspended rootobars coming next, just ahead of hanging or suspended rootograms.

plot 15 of Chapter 12

A different format for two of the displays of plot 14





When, as often, we need to fit the comparison curve to data, we lose little if the quality with which the final display shows what we want to see -- what remains unfitted -- if we are a little crude in our fitting. Putting the hinges in about the right place is usually good enough. We are prepared to do this for rectangular, Gaussian, and Cauchy comparison curves. (Section 12T supplements these three with other alternatives.)

Rather often the best basis for comparison is not some theoretico-mathematical curve, but rather the total behavior of several patterns of counts -- patterns which are, in some useful sense, parallel to one another. Here we can use the principles of

- hanging or suspension.
- square root counts (as a minimum).

Here, again, while rootobars would gain the area principle, it is not likely to be worthwhile to try to use them.

Comparison of patterns of counts boils down to a few essentials:

- reasonable wisdom in the choice of the standard of comparison.
- hanging or suspension so that detailed comparisons are relative to a straight line.
- using square roots, or possibly logs, of counts.
- keeping the displays informative yet with minimum confusion.

*****

12M. How closely does the comparison curve need to be fitted?

Close to the start of section 12B we were a little arbitrary. The quartiles of our 120 residuals were so placed that the intervals corresponding to the units digit of the residuals were about 0.59 long, and the boundary between the center two intervals came at about +0.06, all in units of the half midspreads. We replaced 0.59 by 0.60 and 0.06 by 0.00 with very little discussion.

Did these shifts matter? What size shifts can we safely make -- for pure convenience -- in other situations? How are we to notice that the fit we are using needs adjustment?

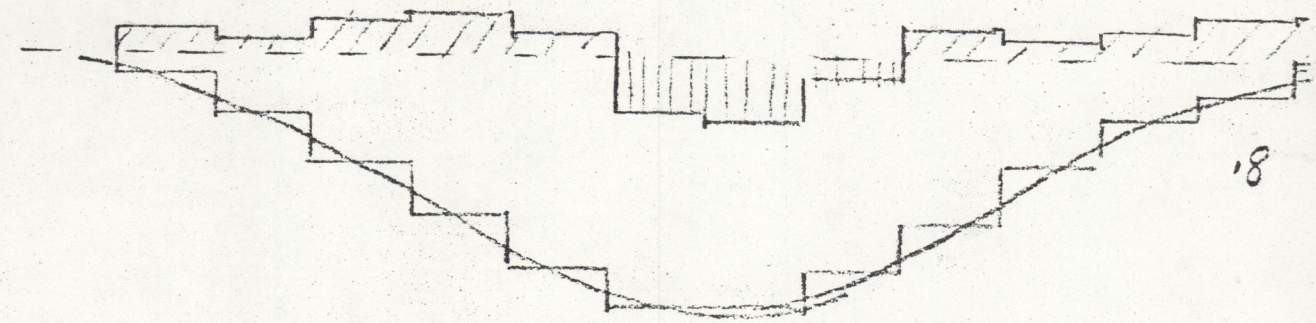
The most direct approach to these questions seems to be the best one. Let us take our batch of 120 residuals, change the comparison curves, and see what effect this has on the resulting pictures. Since we should do this for the most "hit-'em-in-the-eye" presentation that we can find, we will do it for suspended rootograms.

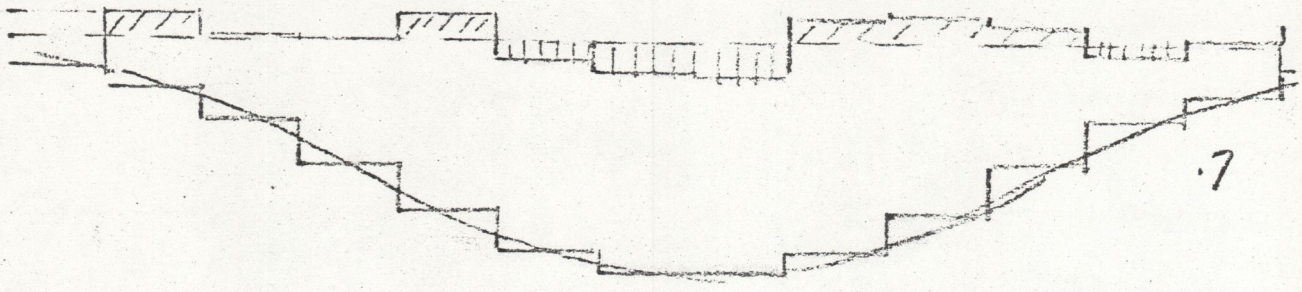
Plot 16 shows the 120 residuals compared with comparison curves of four spreads, corresponding to values of 0.8, 0.7, 0.6 and 0.5 for

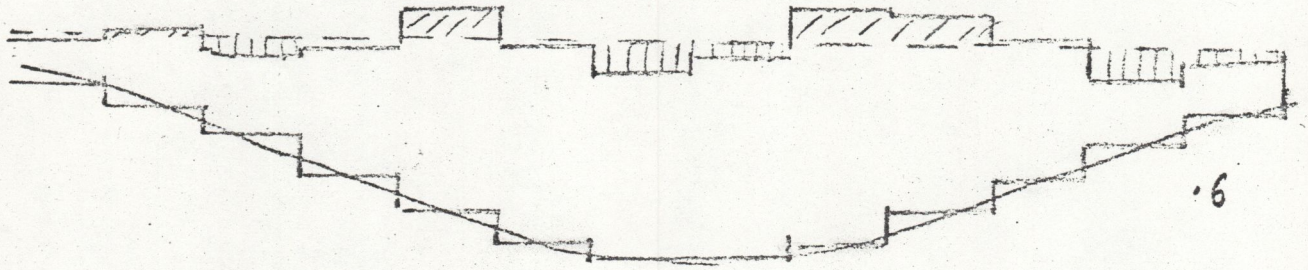
$$\frac{\text{length of each interval}}{\text{half midspread}}$$

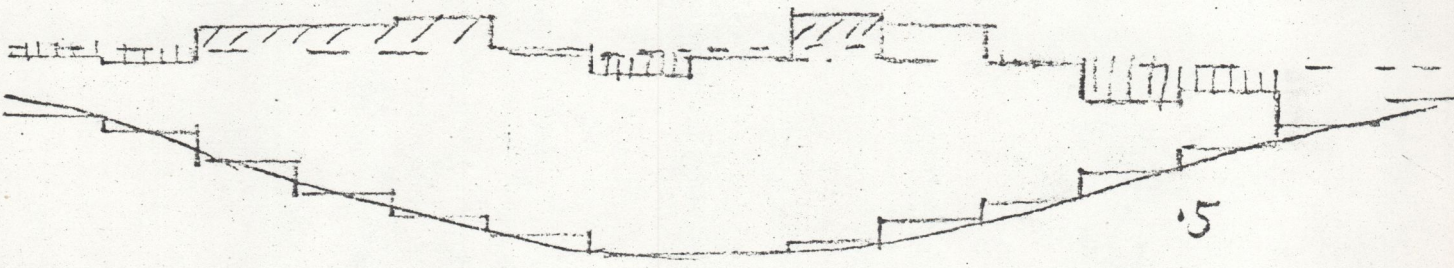
As this ratio goes down, the number of intervals need to cover a given part of the comparison curve goes up. The change from each to the next is detectable, but requires fairly close attention. Overall, as we move down the page, residuals at the ends move from positive to negative, while those in the middle move in the opposite direction. As we

The same observations compared with Gaussian comparison curves of different spread









The values on each suspended rectogram are the ratios assumed for length of intervals in half milliseconds

ought to have expected, a poor fit in terms of spread is represented by a general bowing -- one way in the middle the other at the ends.

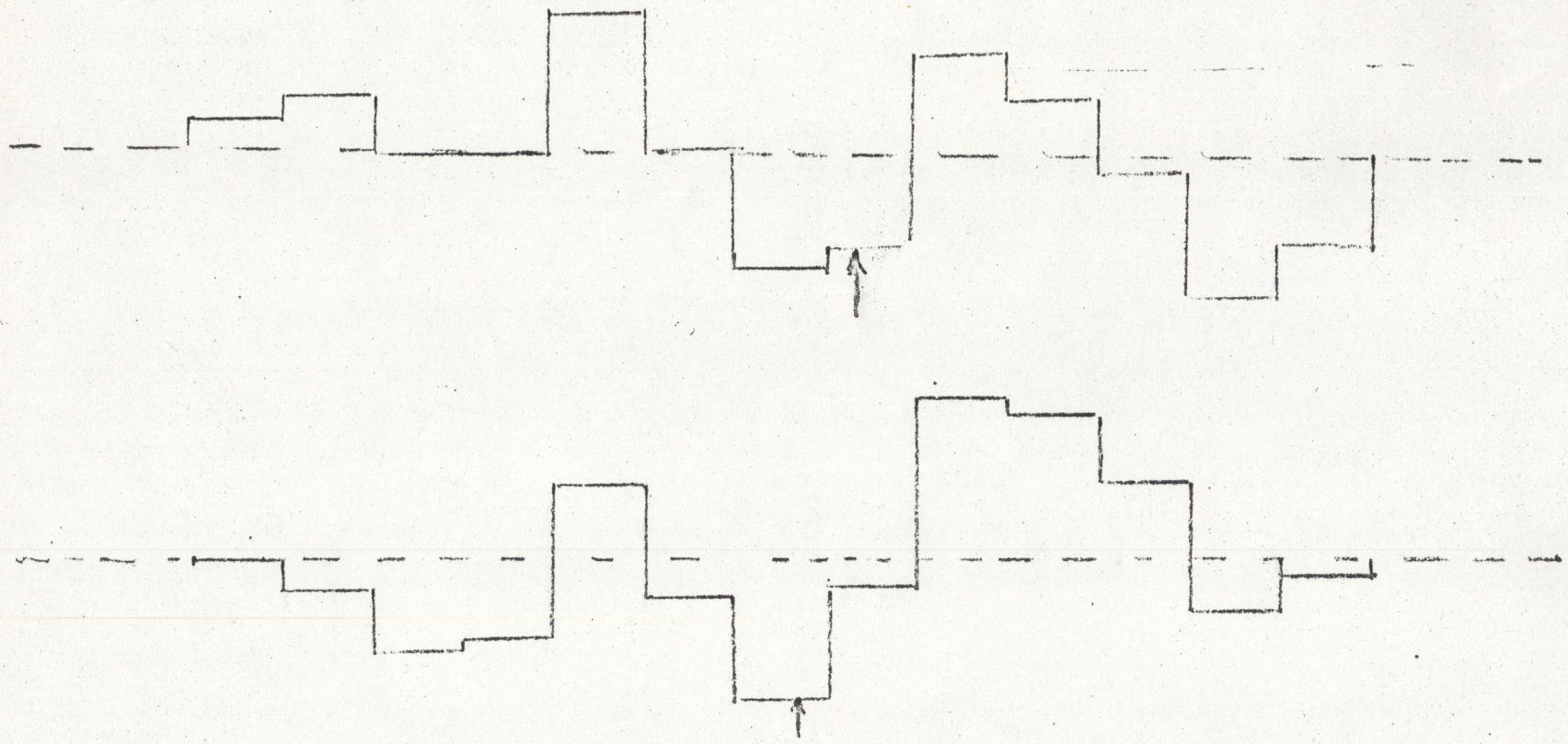
We can feel reasonably sure that neither the fit for .5 nor that for .8 ought to be accepted as satisfactory. Those for .7 and .6 may well be reasonable. With this much uncertainty remaining, the change from .59 to .6 was not very meaningful and hardly deserves comment.

Plot 17 shows the surplus-or-deficit part of the two rootograms corresponding to shifting the center of the comparison curve to the right or to the left by 0.2 midspread. We see that such a shift does produce a detectable tilt in the plot, but does not alter the pattern appreciably.

If our intent is to look hard at the pattern of a set of residuals, making the implicit reference at least nearly horizontal is helpful. But, so far as pattern-looking goes, a nearly horizontal implicit reference is so nearly as good as a quite horizontal one as to make great care in making the implicit reference horizontal far from worthwhile. This is a general principle, and we should not be surprised to see it working here too.

Shifts of ±0.2 were detectable as shifts, but did nothing to distort the picture. Taking +0.06 as 0.00 was thus a wholly negligible change.

plot 17 of chapter 12



Effect on deficit and surplus of shifting the comparison curve sideways by ±0.2 in units of the half midspread. (Arrows show location of center of comparison curve.)

*****

12P. IEDA processor features

The first version of the IEDA processor has no special features adapted to the techniques of this chapter.

12R. Linguistic distributions

We had an example in the last chapter -- table 22 of chapter 11 -- of a distribution of counts based on written language. The comments that follow would, in many ways, have been appropriate there. They were held back to this section for one reason: they mention the Gaussian distribution, which we had not needed until the middle of the present chapter.

Various aspects of the use of words are naturally studied by counting, and by looking at distributions of counts. Among statisticians at least, the study of distributions of sentence lengths was begun by Yule (1939, see also his book of 1944). In the same year, Williams pointed out the closeness of distributions of logs of sentence lengths to the Gaussian shape. (In this connection, Wake's paper (1957) on sentence-length distributions of Greek authors deserves attention.)

The point for us is not that logarithms of sentence lengths are distributed exactly in a Gaussian way, for they are not. (As Wake says, p. 333: "If these distributions are transformed to a logarithmic scale we find that those derived from Aristotle have their upper tails missing whereas for Plato they approximate more nearly to a normal curve in the tail region.") The point is, as Wake's words illustrate, that the actual distributions of the logarithms are close enough to Gaussian to make comparison with a Gaussian curve useful.

That this should be so for sentence length should not surprise us. Sentences of fewer than four or five words

are very rare in connected prose, so that lengths of 4, 3, 2 and 1 will be successively less frequent, not more. Thus the distribution of sentence length is not J-shaped. At the other tail, sentences of much more than 100 words are quite rare -- that tail is long -- when we look at the raw count -- but not extremely long. As an example of how logs help again, sentence lengths are very helpful, but we can -- and will -- leave details for problems (see 12Y).

The frequencies of use of different words are of a quite different character. In a large corpus made up of connected prose, a very common word can occur many, many times. At the other tail, many words will appear only three times each, more twice each, still more once each. Indeed, of the 80 thousand words -- more or less -- to be found in an unabridged English dictionary, the vast majority will be absent. Thus there will be even more words with 0 occurrences than with one. At both tails the distribution of word frequencies is J-shaped. As a result, taking logs is not nearly so likely to bring these distributions under control. Here the plot of (sum of ranks)x(sum of sizes) against log sum of ranks, as illustrated in section 11F, is likely both to be needed and to be effective

12S. <u>What to do when mid-ordinates are not precise enough</u>

We have planned to take the height of curve at the middle of the interval as proportional to the corresponding area. This is usually good enough, but sometimes we have to deal with very long intervals. In such a situation use a formula derived from what is often called Simpson's Rule, namely:

- take the area proportional to:

(height at one end of interval)
+ 4(height in middle)
+ (height at other end)

(We could divide by 6, but, for our use, the scaling factor will take care of this at no extra cost.)

12T. Fitting comparison curves related to folded powers

Every form of expressing fractions corresponds to a shape of distribution. For we can start with a rectangular distribution extending from 0 to 1, and say:

- if the raw fraction has a rectangular (uniform) distribution, what then is the distribution of our chosen expression?

Doing this with folded powers -- including folded logs as a special case, produces an interesting and useful set of distributions. Those that are most easily manageable are those corresponding to powers, 1, 1/2, 0, -1/2, and -1. Fortunately these cover a quite wide range of symmetrical shapes, and come close enough to one another so that each of a wide variety of purposes can be served by someone.

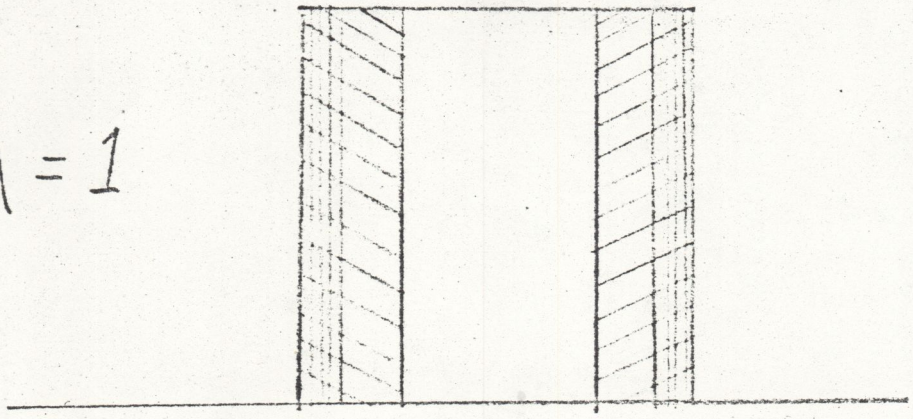
Just as each of these changes of expression stretched out the extremes, so now these new distributions will be more stretched out, and their comparison curves lower, toward their ends.

Distributions of the shape arising for $\lambda = 0$ have been used for a long time under the name "logistic". By analogy, then, we may use "rootistic" for $\lambda = 1/2$, "tooristic" for $\lambda = -1/2$ and "invistic" for $\lambda = -1$. Here we have turned "root" around as "toor" to indicate the change from a positive to a negative power. The shapes of these distributions are shown in plots 18 and 19, where the quartile positions and central heights have been matched.

As λ decreases, we can see the tails growing longer and longer -- and longer. With the quartiles in fixed position,

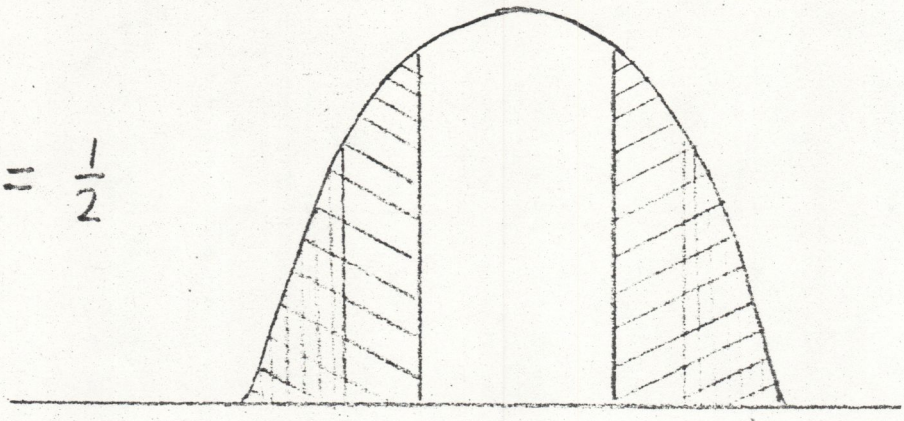
Some distributions related to folded powers

$\lambda = 1$



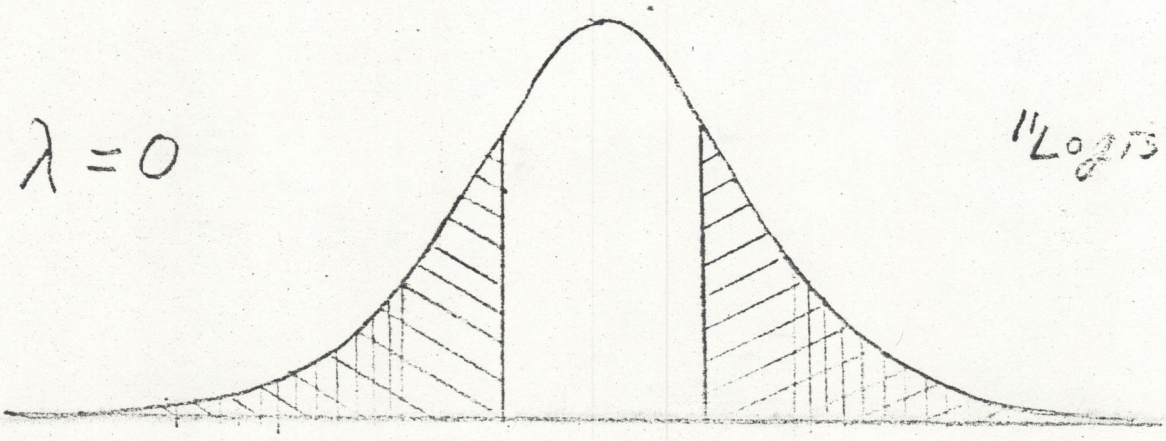
"Rectangular"

$\lambda = \frac{1}{2}$



"Rootistic"

$\lambda = 0$

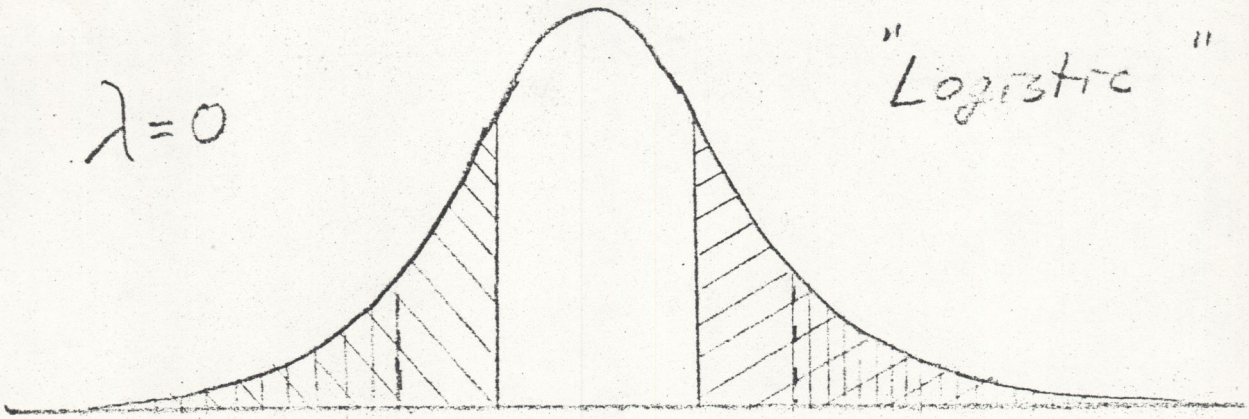


"Logistic"

Note: Middle 50% unshaded, both end 10%'s darkly shaded.

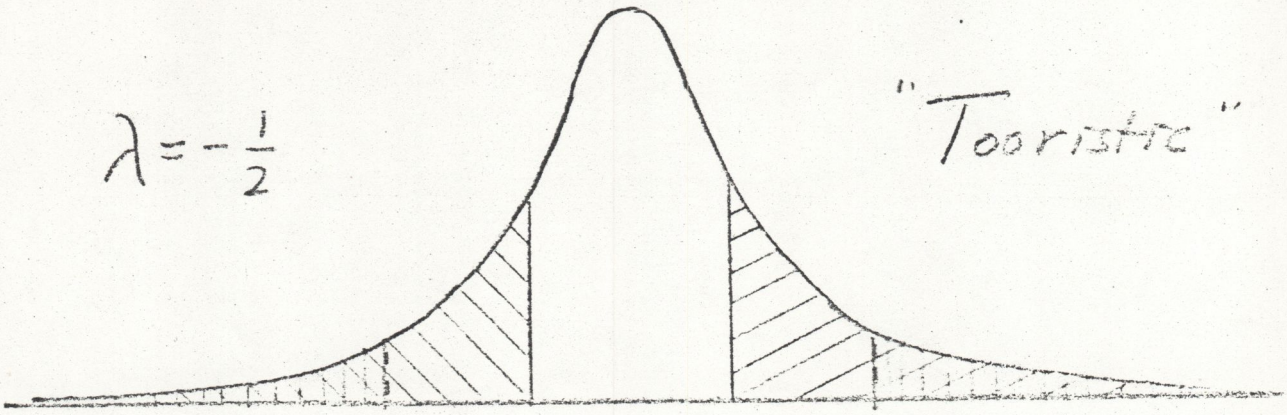
More distributions related to folded powers

$\lambda = 0$



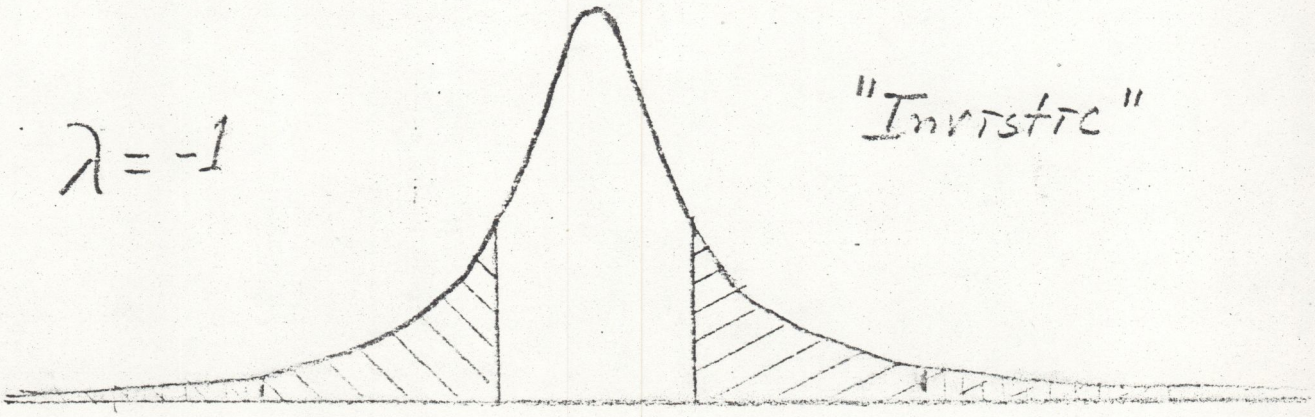
"Logistic"

$\lambda = -\frac{1}{2}$



"Tooristic"

$\lambda = -1$



"Inristic"

Note: Middle 50% unshaded, both end 10%'s doubly shaded.

the points cutting off 10% on each tail, shown by the change from single to double shading, move further and further out as λ moves from +1 through 0 to -1.

It should not surprise us that this connection between distributions, on the one hand, and expressions of fractions, on the other, is a two-way street. Not only can we get distributions from expressions of fractions -- we can also -- as people have long done -- get expressions for fractions from distributions. It is from the Gaussian distribution that normits and probits -- mentioned in 10M as expressions for fractions -- were gotten.

Since normits and probits can be closely approximated by folded powers with $\lambda = 0.14$, it follows that the distribution of the sort we are here considering for $\lambda = 0.14$ is a close approximation to the Gaussian distribution. It happens to be true that the distribution for $\lambda = -0.97$, which is close to the invistic distribution, is a good approximation to the Cauchy distribution, which we took in section 12C as a fine example of a long-tailed distribution. Thus the family of λ-distributions, with selected instances from which we have been concerned, in this section, comes close to the Gaussian and the Cauchy distributions, as well as including the rectangular.

Table 20 summarizes the behavior of the λ-distributions for $\lambda = 0.5$, 0, -0.5 and -1 on a scale based on the half midspread, both as to height of curve and as to remaining fraction of area.

For those who need more detail (in particular for the fitting of comparison curves in the style of 12C), table 21 gives more detailed numerical values.

Summary Behavior of λ-distributions

	Relative height of curve						Fraction of area beyond x					
x	λ=1	0.5	λ=0	-0.5	λ=-1	x	λ=1	0.5	λ=0	-0.5	λ=-1	x
0.0	1.00	1.00	1.00	1.00	1.00	0.0	.500	.500	.500	.500	.500	0.0
0.5	1.00	.97	.93	.85	.75	0.5	.375	.372	.366	.352	.351	0.5
1.0	1.00	.90	.75	.59	.45	1.0	.250	.250	.250	.250	.250	1.0
1.5	1.00	.76	.54	.39	.27	1.5	.125	.142	.161	.178	.19	1.5
2.0	cuts	.54	.36	.25	.18	2.0	.000	.057	.110	.131	.153	2.0
2.5	.00	.21	.23	.17	.13	2.5	"	.007	.060	.100	.128	2.5
3.		.00	.14	.12	.09	3.		none	.036	.078	.110	3.
4.		"	.05	.07	.06	4.		"	.012	.051	.086	4.
5.		"	.016	.04	.04	5.		"	.004	.036	.069	5.
6.		"	.006	.02	.03	6.		"	.001	.027	.059	6.
7.		"	.002	.017	.02	7.		"	.000	.021	.051	7.
8.		"	.001	.012	.016	8.		"	"	.017	.045	8.
10.			.000	.006	.010	10.				.011	.036	10.
15.			"	-	.005	15.				-	.025	15.
20.			"	-	.002	20.				-	.019	20.

Notes: (1) x is in units of the half midspread.
(2) some values for λ = -1.0 are interpolated.

table 21 of chapter 12

Information on λ-distributions Suitable for Fitting Comparison Curves (Entries are relative heights, followed in parentheses by remaining fractions)

x	λ=0.5	λ=0	λ=-0.5	λ=-1	x
0.0	1.00 (.500)	1.00 (.500)	1.00 (.500)	1.00 (.500)	0.0
0.1	1.00 (.476)	1.00 (.473)	.99 (.470)	-	0.1
0.2	.99 (.448)	.99 (.445)	.97 (.441)	.95 (.434)	0.2
0.3	.99 (.423)	.97 (.418)	.94 (.413)	-	0.3
0.4	.98 (.397)	.95 (.392)	.90 (.384)	.83 (.375)	0.4
0.5	.97 (.372)	.93 (.366)	.85 (.358)	-	0.5
0.6	.96 (.347)	.90 (.341)	.80 (.333)	.68 (.325)	0.6
0.7	.95 (.322)	.87 (.317)	.75 (.310	-	0.7
0.8	.93 (.297)	.83 (.293)	.70 (.289)	.55 (.283)	0.8
0.9	.92 (.273)	.79 (.271)	.64 (.268)	-	0.9
1.0	.90 (.250)	.75 (.250)	.59 (.250)	.45 (.250)	1.0
1.1	.87 (.227)	.71 (.230)	.55 (.233)	-	1.1
1.2	.85 (.205)	.67 (.211)	.50 (.217)	.37 (.223)	1.2
1.3	.82 (.183)	.62 (.293)	.46 (.203)	-	1.3
1.4	.79 (.162)	.58 (.177)	.42 (.190)	.30 (.201)	1.4
1.5	.76 (.132)	.54 (.161)	.39 (.178)	-	1.5
1.6	.72 (.123)	.50 (.147)	.35 (.167)	.25 (.182)	1.6
1.7	.68 (.105)	.46 (.134)	.32 (.157)	-	1.7
1.8	.64 (.088)	.43 (.122)	.30 (.148)	.21 (1.67)	1.8
1.9	.59 (.072)	.39 (.110)	.27 (.139)	-	1.9
2.0	.54 (.057	.36 (.100)	.25 (.131)	.18 (.154)	2.0
2.1	.49 (.044)	.33 (.091)	.23 (.124)	-	2.1
2.2	.43 (.032)	.30 (.082)	.22 (.117)	.16 (.142)	2.2
2.3	.36 (.022)	.27 (.074)	.20 (.118)	-	2.3
2.4	.29 (.013)	.25 (.067)	.19 (.105)	.14 (.132)	2.4
2.5	.21 (.007)	.23 (.060)	.17 (.100)	-	2.5
2.6	.13 (.002)	.20 (.054)	.16 (.095)	.12 (.124)	2.6
2.7	.03 (.000)	.19 (.049)	.15 (.090)	-	2.7
2.8	.00 (.000)	.17 (.045)	.14 (.086)	.11 (.116)	2.8
2.9	"	.15 (.040)	.13 (.082)	-	2.9
3.0	"	.14 (.036)	.12 (.078)	.09 (.110)	3.0

(Cont'd)

table 21 (Cont'd)

x	λ=0	λ=-0.5	λ=-1	
3.0	.137 (.036)	.120 (.078)	.094 (.110)	3.0
3.2	.112 (.029)	.106 (.071)	.085 (.104)	3.2
3.4	.090 (.023)	.093 (.065)	.076 (.098)	3.4
3.6	.073 (.0188)	.082 (.060)	.069 (.093)	3.6
3.8	.060 (.0151)	.073 (.056)	.062 (.089)	3.8
4.0	.048 (.0112)	.065 (.051)	.057 (.085)	4.0
4.2	.039 (.0110)	.058 (.048)	.053 (.081)	4.2
4.4	.031 (.0079)	.052 (.044)	.048 (.078)	4.4
4.6	.025 (.0063)	.047 (.041)	.046 (.075)	4.6
4.8	.020 (.0051)	.043 (.039)	.041 (.072)	4.8
5.0	.016 (.0041)	.039 (.036)	.038 (.069)	5.0
5.5	.009 (.0024)	.031 (.031)	-	5.5
6.0	.006 (.0014)	.025 (.027)	.027 (.058)	6.0
6.5	.003 (.0008)	.020 (.024)	-	6.5
7.0	.001 (.0005)	.017 (.021)	-	7.0
7.5	.001 (.0003)	.014 (.019)	-	7.5
8.0	.000 (.0002)	.012 (.017)	.016 (.045)	8.0
8.5	" (.0001)	.010 (.015)	-	8.5
9.0	" (.0001)	.009 (.013)	-	9.0
9.5	" (.0000)	.007 (.012)	-	9.5
10.0	" (.0000)	.006 (.011)	.010 (.036)	10.0

λ = 0 is "logistic".

λ = -0.5 is "tooristic".

λ = -1 is "invistic".

*****

12Y. Problems

1. Turning back to chapter 11, table 22 of chapter 11 gave the distribution of frequencies of occurrence of words in Pushkin's The Captain's Daughter. The distribution of "logs" of number of occurrences was treated in problem 4 of section 11Y. Fit a Gaussian comparison curve to this distribution of logs, and compare the observed distribution to the curve pictorially.

2. Williams (1939) reported the counts of lengths of 600 sentences each from:

G. K. Chesterton: A Short History of England 1917.

G. Bernard Shaw: An intelligent woman's guide to Socialism.

The results are summarized in table 22. Williams gave (his page 359) plots showing the effectiveness of comparison of the distribution of logarithms of sentence lengths with Gaussian curves. (This is the method of choice for analyzing this data.)

Williams remarked (his page 360) that: "It is also perhaps worthy of passing comment that the curve representing Mr. Shaw is short and broad while that representing Mr. Chesterton is tall and slender; which shows how necessary it is to use these curves only for the purpose for which they were originally designed." Make rank-size-log rank plots for these two distributions in order to see if:

- this (less effective) mode of plotting does manage to contrast these two authors' styles effectively.

- if Williams's remarks apply as well to this plot as to the distribution of logarithms of sentence lengths.

table 22 of chapter 12

Summaries of Williams's Counts of 600 Sentence Lengths

G. K. Chesterton				G. Bernard Shaw			
Length	Count	Length	Count	Length	Count	Length	Count
91	1	29	21	143	1	29	13
76	1	28	22	141	1	28	14
60	1	27	21	115	1	27	13
59	1	26	28	111	1	26	10
58	2	25	19	107	2	25	13
56	2	24	22	99	1	24	13
55	2	23	19	96	3	23	11
53	2	22	26	90	2	22	18
52	2	21	21	87	2	21	12
51	3	20	25	86	1	20	16
50	8	19	24	84	2	19	12
49	2	18	28	83	3	18	22
48	2	17	20	82	1	17	16
47	3	16	16	81	3	16	12
46	4	15	15	80	1	15	21
45	4	14	16	75-79	7	14	16
44	4	13	16	70-74	8	13	14
43	8	12	17	65-67	6	12	6
42	6	11	7	60-64	10	11	9
41	6	10	6	55-59	15	10	15
40	10	9	11	50-54	38	9	16
38-39	13	8	2	45-49	25	8	11
36-37	18	7	5	40-44	28	7	10
34-35	14	6	3	36-39	25	6	10
32-33	29	5	3	34-35	23	5	14
30-31	38			32-33	18	4	3
Total	599			30-31	30	3	2
				Total	600		

Source: C. B. Williams 1939.